Scientific and Technical Report

Sponsored by
Advanced Research Projects Agency/ITO

Browsing, Discovery and Search in Large Distributed Databases
of Complex and Scanned Documents

ARPA Order No. D468

Issued by EXC/AXS under Contract #F19628-95-C-0235

Date Submitted: October 10, 1996

Period of Report: June 26, 1996 to September 30, 1996

Submitted by: Professor W. Bruce Croft, Principal Investigator
Computer Science Department
University of Massachusetts, Amherst

Distribution Statement A: Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	10/10/96	Scientific/Tech 6/26/96 - 9/30/96

4. TITLE AND SUBTITLE
Browsing, Discovery, and Search in Large Distributed Databases of Complex and Scanned Documents

5. FUNDING NUMBERS
F19628-95-C-0235
ARPA Order No. D468

6. AUTHOR(S)
W. Bruce Croft

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Massachusetts, Amherst
Box 36010, OGCA, Munson Hall
Amherst, MA 01003-6010

8. PERFORMING ORGANIZATION REPORT NUMBER
TR5281811096

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Mr. Harry Koch
ESC/AXS
Bldg 1704. Room 114
5 Eglin St.
Hanscom AFB, MA 01731-2116

Ms. Monique Dillon
Office of Naval Research
Boston Regional Office
495 Summer St., Room 103
Boston, MA 02210-2109

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Statement A: Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This project aims to integrate powerful, new techniques for interactive browsing, discovery, and retrieval in very large, distributed databases of complex and scanned documents. Emphasis is placed on going beyond full-text retrieval techniques developed in the DARPA TIPSTER program to support different types of access and non-textual content. These techniques should be particularly relevant to the patent domain where it is important to find relationships between documents and where the patent or trademark may be based on a visual design. The specific tasks identified involve studying representation techniques for long documents with complex structure, browsing and discovery techniques for large text databases, image retrieval and scanned document retrieval techniques, and architectures for large, distributed databases.

14. SUBJECT TERMS

Browsing	Query Processing	Indexing
Image Retrieval	Scanned Document Retrieval	Bayesian Network
Text Retrieval	Probabilistic Retrieval Model	Large Distributed Databases

15. NUMBER OF PAGES
10

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Browsing, Discovery and Search in Large Distributed Databases of Complex and Scanned Documents

Technical and Scientific Report

Task 1: Representation Techniques for Complex Documents

Task Objectives

In this task, the goal is to extend the word-based representations that are common in retrieval systems in order to support summarization, browsing, and more effective retrieval. Specifically, we will be studying phrase-based representations and relationships between phrases in individual and groups of documents as the basis for our approach. Document structure will be used as part of the information that is used to "tag" the phrasal representation.

Technical Problems

The technical problems have to do with defining a "phrase", developing techniques for rapidly extracting them from text, comparing phrase contexts to identify significant relationships, producing summaries from these representations, and extending the underlying retrieval model to be able to make effective use of phrasal representations in both query-based retrieval and relevance feedback.

General Methodology

The general methodology for this task is to demonstrate effectiveness through user-based and collection-based experiments. Extensive use will be made of the TIPSTER document collection, which consists of a large number of text documents from a variety of sources, queries, and user relevance judgments for each query. This collection will be used for the experiments involving new probabilistic retrieval models and relevance feedback. Summarization techniques will be compared to sentence-based approaches and user-based evaluations of these summaries will be done. As more work is done on summarization in the TIPSTER program, we will make use of any new evaluation measures developed there.

Technical Results

The initial period of the project has involved getting access to patent databases and acquiring appropriate hardware for the experiments. We have begun to look at phrase-

based indexing and retrieval in three specific ways; first, we have extracted all significant phrases from the entire TIPSTER database and are examining this database to determine which of these should be used for indexing. Second, we have been working on techniques to automatically identify the most significant phrases and word relationships in a query and use this to contract more effective models. Third, we are continuing to experiment with the use of phrases and phrase-like patterns in machine learning approaches using relevance feedback.

Important Findings and Conclusions

It is too early in the project to claim any conclusive results. We will have some more experimental evidence after the TREC conference in November.

Significant Hardware Development

Purchases were made of disk to store large collections of patent data.

Special Comments

We continue to work with the PTO, San Diego Supercomputer consortium and DARPA to obtain access to some of the patent collections and establish fast network links in order to be able to use the very large archives of scanned patents.

Implication for Further Research

Plan is to continue experiments on phrase-based indexing and extensions to probabilistic model. As soon as patent data is available, phrase experiments will be done with this.

Task 2: Browsing and Discovery Techniques for Document Collections

Task Objectives

The goals of this task are to develop techniques for summarizing collections of documents, and discovering connections between important ideas and documents in distributed collections. These techniques will be designed to support interactive browsing in environments like the PTO.

Technical Problems

The technical problems involve producing an effective summary of a group of documents, such as a retrieved set or an entire database. Both document and phrase clusters could be used as part of this process. In order to support discovery, connections must be made

between documents and groups of phrases that use a variety of evidence in addition to direct co-occurrence.

General Methodology

The techniques will be evaluated with user-based and collection-based experiments. The relevance judgments from the TIPSTER collection will be used to evaluate clusters of documents. Phrase clusters will be evaluated by their impact on retrieval effectiveness and through user experiments that will measure performance on specific tasks. Part of the effort in this task (and the previous one) will involve developing a PTO test collection, which means that sample queries will need to be gathered from patent examiners and they will need to evaluate demonstrations of tools as they are developed.

Technical Results

Some initial work has been done with phrase and document clustering of collections and retrieved sets. We have developed a first demonstration of a 3-D graphics interface designed for manipulating document and concept relationships to identify strong groupings and relationships. The first evaluation of this work was a simple user experiment to test feasibility. This work is described in a technical report.

Important Findings and Conclusions

Our initial finding is that visualizations of document and concept relationships can be very useful in helping a searcher more accurately locate relevant material. This initial finding needs to be supported by more extensive experiments that have just got underway.

Significant Hardware Development

None.

Special Comments

We continue to work with the PTO, San Diego Supercomputer consortium and DARPA to obtain access to some of the patent collections and establish fast network links in order to be able to use the very large archives of scanned patents.

Implication for Further Research

Plan is to continue developing techniques for browsing in document and concept space and to scale up experiments.

Task 3: Scanned Document Indexing and Retrieval

Task Objectives

The goals of this task are to develop techniques for detecting text, trademarks, logos, and images in scanned documents, clean up backgrounds of these detected objects, and support retrieval of images (such as designs in design patents), trademarks, and text from OCR.

Technical Problems

Current zoning techniques available with commercial OCR devices do not accurately locate text or trademarks within other images. We are developing techniques based on gaussian derivative filters to both detect and clean up (remove noisy backgrounds) these classes of objects in scanned documents. We are developing "appearance-based" retrieval of images as well as more straightforward features such as color and texture. Filter based and frequency domain based techniques offer some potential in this area, but significant work needs to be done on making this approach efficient enough to deal with thousands of images.

General Methodology

The evaluation of these techniques will be done in a similar way to text by developing test collections of images and scanned documents. Specifically, we are working to obtain large collections of trademarks and design patents, as well as typical queries.

Technical Results

We have carried out a number of experiments with smaller databases on techniques for text identification and image retrieval. This work is described in technical reports. We have also made some progress on the problem of indexing images to improve retrieval efficiency, but much remains to be done.

Important Findings and Conclusions

Initial results suggest that appearance based retrieval of images is feasible and can produce good results. More extensive experiments using patent data need to be done. Initial results with text detection are also promising.

Significant Hardware Development

Disk acquisition.

Special Comments

Gaining access to PTO images has been a priority.

Implication for Further Research

Plan is to scale up image retrieval experiments to test indexing impact on retrieval speed. Also must start to acquire typical queries from patent examiners.

Task 4: Distributed Retrieval Architecture

Task Objectives

The goals of this task are to scale up our current methods of automatically selecting collections and merging results, and to investigate architectures that can support efficient retrieval, browsing and relevance feedback in distributed environments with terabytes of information.

Technical Problems

The current INQUERY text retrieval system uses a client server architecture to support simultaneous retrieval from multiple collections distributed across one or more processors. A number of efficiency bottlenecks develop, however, when the size of the databases is very large. Deciding which subcollections to search can address part of the problem, but there are other problems associated with the fundamental efficiency of the processes involved and the use of distributed resources. Image indexing and retrieval tends to make all of these problems worse since the databases and indexes are considerably larger.

General Methodology

The architectures and algorithms produced in this task will be evaluated using a combination of standard performance (efficiency) measures and effectiveness measures. The efficiency tests will be done using large PTO databases, including images, and the collection selection algorithms will be evaluated using the text subcollections of the patents.

Technical Results

We have developed a first version of a multi-threaded INQUERY and are currently comparing the efficiency of this and the previous system in a client-server environment.

A first version of collection selection and result merging has been incorporated into INQUERY.

Important Findings and Conclusions

Too early for any conclusions.

Significant Hardware Development

Purchases were made of disk to store large collections of patent data.

Special Comments

Previous comments on fast network access to other PTO sites are particularly relevant here, since this will be required to both test the distributed architecture and to index and retrieve the full versions of the PTO databases.

Implications for Further Research

Plan is to begin more experiments on collection selection and result merging as well as continuing performance experiments.

Distribution List

ARPA Agent: Harry Koch
ESC/AXS
Bldg 1704, Rm 114
5 Eglin Street
Hanscom AFB, MA 01831-2116

ARPA/ITO
ATTN: Gary Koob
3701 N Fairfax Drive
Arlington, VA 22203-1714

ARPA/Technical Library
3701 N Fairfax Drive
Arlington, VA 22203-6145

Defense Technical Information Center (DTIC)
Cameron Station
Alexandria, VA 22034-6145

ESC/ENK
ATTN: Ms Carole Stephan
Bldg 1704, Rm 119
5 Eglin Street
Hamscom AFB, MA 01731-2116
(Letter of Transmittal Only)